J Geograph Syst (1999) 1:219–236

Journal of Geographical Systems

Identifying local spatial association in flow data*

Svante Berglund, Anders Karlström

Department of Infrastructure and Planning, Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: svante@infra.kth.se)

Received: 18 February 1998 / Accepted: 29 September 1998

Abstract. In this paper we develop a spatial association statistic for flow data by generalizing the statistic of Getis-Ord, G_i (and G_i*). This local measure of spatial association, G_{ij}, is associated with each origin-destination pair. We define spatial weight matrices with different metrics in flow space. These spatial weight matrices focus on different aspects of local spatial association. We also define measures which control for generation or attraction non-stationarity. The measures are implemented to examine the spatial association of residuals from two different models. Using the permutation approach, significance bounds are computed for each statistic. In contrast to the G_i statistic, the normal approximation is often appropriate, but the statistics are still correlated. Small sample properties are also briefly discussed.

Key words: Local spatial association, GIS, flow data

JEL classification: C14, R12, R41

1 Introduction

Recently there has been a growing interest in local indicators of spatial association. This is partly generated by the need and possibility to analyse the large amount of spatial data that has become available. Anselin and Bao (1997) argues that most of the global measures of spatial association were developed when computer power was scarce and data sets were small. Global measures of spatial association and spatial autocorrelation do not take advantage of the capabilities of GIS. The local indicators of spatial association, on the other hand, is more informative in that they can be visualised in a GIS, and pockets of nonstationarities (hot spots and spatial outliers) can be identi-

* An earlier version of this paper was presented at the European Regional Science Association Conference in Rome, Italy, August 1997. Part of this research has been financed by financial grants from the Swedish Transport and Communications Research Board and Center for Geoinformatics, Royal Institute of Technology, Sweden.

fied. One intuitive measure of local spatial association that has gained wide spread acceptance is the G_i statistic, which was introduced by Getis and Ord (1992). The G_i statistic identifies neighbourhoods of zones with high or low values. This measure of local spatial association was generalised in Ord and Getis (1995) to allow for non-binary spatial weight matrices and non-positive values. This generalisation allows the G_i statistic to be applied to residuals from a model, which provides us with a tool for assessing the structure of the error term and identifying nonstationarities. Although the G_i statistic is formulated for data with scalar values, it can be generalised to vector data. The G_{ij} statistic suggested in this paper will be associated with each origin and destination pair.

In flow models, such as spatial interaction models of migration and traffic, spatial association and network association are important aspects. However, the spatial structure of the error term in flow models has not until recently received much interest in the literature[1]. Black (1992) distinguishes between network autocorrelation and spatial autocorrelation. *Spatial autocorrelation* in his terminology concerns the influence of variables in one location on variables in other (neighbourhood) locations. Network autocorrelation, on the other hand, concerns the influence on values associated with a link on other links which are interconnected. Black (1992) examines *global* network autocorrelation using Moran's I. The objective of this paper, on the other hand, is to identify *local* network and spatial association with the use of the G_i statistic.

The G_{ij} statistic proposed in this paper is applied to measure local spatial association in residuals from flow-models. In the context of the modelling process, measures of this kind apply to the model evaluation. In model evaluation, tools for assessment of the structure of the error term are important. For some estimators, underlying assumptions of the distribution may be violated and corrections may be necessary. The tools at hand for testing and evaluating models in a spatial sense can be divided into two categories, focused and general tests, see Besag and Newell (1991). General tests, e.g. Moran's I and Geary's c, return a number (global statistic) similar to the Durbin-Watson statistic in time series analysis. Hence, general tests give us information whether spatial correlation is present or not. They give no answer whether the correlation is more pronounced at certain locations. This is a weakness, but in return general tests are robust in a statistical sense due to the often large number of observations at hand. Once a general test statistic is calculated, space is eliminated. The preservation of space in focused tests gives a more descriptive character. A well known focused test statistic (or localised statistic) is the G_i (and G_i^*) statistic mentioned above. Localised statistics can be mapped, and can reveal nonstationarity that global statistics fail to identify. Fotheringham (1994) argues that localised statistics are more informative, if they can reveal how relationships between different sets of variables vary over space. Although the local G_i statistic can identify *where* there are nonstationarities, it can not by itself reveal the nature of how relationships between variables vary over space. We show that using different spatial weight matrices we can gain information of the nature of nonstationarity.

[1] Bolduc (1992) and Bolduc et al. (1995) are notable exceptions. In a model of mode choice, they split the error component in three parts, one part related to origin zones, one related to destination zones, and one part related to the link from origin to destination. Estimating this model gives evidence of global autocorrelation.

In this paper we analyse spatial association in flow patterns, by applying the G_{ij} statistic to residual flows from two models, one migration model and one commuting model. The two models are quite different and demonstrate the usefulness of the G_{ij} statistic as a local measure of spatial association. The theoretical property of normally distributed residuals from a properly estimated OLS allows us to make comparisons between the conditional permutations approach and normal approximations in the first model. The second model enables us to address small sample properties of the proposed statistic and methods for establishing significance bounds when the underlying distribution is unknown.

The paper is organised as follows. In section two we define the statistic in its general form, and define different measures to be used in the applications. In section three the results of the first application are reported. In this application, the structure of the error term in an OLS migration model is assessed, and more specifically local spatial nonstationarities are identified. The conditional permutations approach is employed for (pseudo-)significance tests, and its properties are discussed. In section four we employ the G_{ij} statistic to analyse residuals from a logit model of commuting patterns in a small scale application. Distribution and correlation properties of the measures are discussed. In the concluding section we discuss the results.

2 Definition of the G_{ij} statistic

As a starting point we take the G_i statistic (Getis and Ord 1992), which is simply taken as the ratio of sum of values in a neighbourhood around a location (zone) to the sum of all values (in the whole sample), as follows

$$G_i = \frac{\sum_{j,j\neq i} w_{ij} y_j}{\sum_{j,j\neq i} y_j}, \tag{1}$$

where w_{ij} is the common binary spatial weight matrix. A G_i^* statistic can also be defined, where the asterisk denotes that we also include the observation at location i. Significance bounds for G_i (and G_i^*) can easily be established if the underlying distribution is normal, and the result can be mapped for visualisation and analysis. A high and significant value in a location indicates spatial clustering of high values. A small and significant value indicates clustering of small values around the location. This is different than the interpretation of many other statistical measures, e.g. Moran's I, where a high value indicates spatial association of similar values, while a small (negative) value indicates spatial clustering of dissimilar values.

In Ord and Getis (1995), (1) was generalized to allow for nonpositive observations as well as nonbinary spatial weights. In its general formulation, given a spatial weight matrix $W = [w_{ij}]$, the statistic is defined as

$$G_i(W) = \frac{\sum_j w_{ij} y_j - W_i \bar{y}}{s\{(n-1)S_1 - W_i^2/(n-2)\}^{1/2}} \tag{2}$$

where

$$\bar{y} = \frac{1}{n-1} \sum_{j, j \neq i} y_j, \tag{3}$$

$$s^2 = \frac{1}{n-2} \sum_{j, j \neq i} (y_j - \bar{y})^2, \tag{4}$$

$$W_i = \sum_j w_{ij}, \tag{5}$$

$$S_1 = \sum_{j, j \neq i} w_{ij}^2, \tag{6}$$

and n equals the number of zones. This generalisation allows the statistic to be applied to residuals from a model.

Although the G_i statistic was defined in the context of scalar observations in each zone, it is easily generalised to flow data. If we, in equation (2), let i denote the flow from i to j, j the flow from k to l, and n the number of flows, equation (2) can be directly applied to flow data. It is more informative to explicitly define the G_i statistic to be applied on flow data. Let r_{ij} denote flows between each pair of zones. Then, given a spatial weight matrix $W = [w_{ij,kl}]$, we define[2]

$$G_{ij}(W) = \frac{\sum_{k,l} w_{ij,kl} r_{kl} - W_{ij} \bar{r}}{s\{(t-1)S_1 - W_{ij}^2/(t-2)\}^{1/2}}, \tag{7}$$

where t equals the number of flows,

$$\bar{r} = \frac{1}{t-1} \sum\nolimits_{kl, (k,l) \neq (i,j)} r_{kl},$$

$$s^2 = \frac{1}{t-2} \sum\nolimits_{kl, (k,l) \neq (i,j)} (r_{kl} - \bar{r})^2,$$

$$W_{ij} = \sum\nolimits_{kl, (k,l) \neq (i,j)} w_{ij,kl},$$

$$S_1 = \sum_{kl, (k,l) \neq (i,j)} w_{ij,kl}^2.$$

In our applications r_{ij} will represent residual flows,

$$r_{ij} = T_{ij} - \hat{T}_{ij}, \tag{8}$$

where T_{ij} is the observed flow between zone i and zone j and $\hat{T}_{ij}$ is the estimated flow by some model.

[2] Including the flow from i to j defines the G_{ij}^* statistic in analogy with the G_i^* statistic.

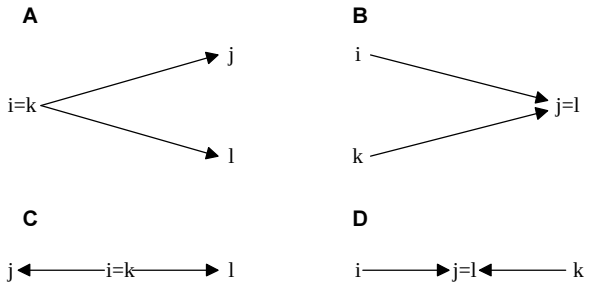

Fig. 1. Illustration of weight matrices

Since equation (7) still is the G_i statistic, the intuition of the measure is still valid when applied to flow data. Computing this statistic for a flow associated with an origin-destination pair (i, j) and establishing significance bounds will indicate that in a neighbourhood of this flow (to be defined below) there is spatial clustering of low or high flows. What remains to be defined is only a metric in flow space, or more specifically, the spatial weight matrix in flow space. Let us consider the simple and often applied binary spatial weight matrix. With scalar data, the meaning of a zone being the neighbour of another zone is clear, but how do we define neighbourhoods of flows?

We will in this paper use two different kind of weight matrices. In the terminology of Black (1992) they focus on different aspects of association, viz. spatial association and network association. Spatial association, by this definition, is association that falls back on the spatial weight matrix among zones, in other words the configuration of zones. We will use the following two binary spatial weight matrices (illustrated in Fig. 1 A and B, respectively):

$$W^d = [w_{ij,kl}] = \begin{cases} 1 & \text{if } i = k \text{ and } w_{jl} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$W^o = [w_{ij,kl}] = \begin{cases} 1 & \text{if } j = l \text{ and } w_{ik} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where w_{ij} denotes elements of the traditional binary spatial weight matrix, i.e. w_{ij} equals one if i and j are neighbours, and zero otherwise. Thus, in the numerator of equation (7), with W^d we sum over flows from i to a neighbourhood of j, and with W^o we sum over flows from a neighbourhood of i to j.

Although a binary spatial weight matrix is often used in applications, it is often seen only as a first approximation, and more general forms of weight matrices should be considered. It is often, however, difficult to single out one particular weight matrix from a number of candidates. Following Bolduc (1992) and Bolduc et al. (1995) we have also employed a more general form, with parameterised spatial weights,

$$W^\theta = [w_{ij,kl}] = (d_{il} + d_{jk})^{-\theta}, \tag{11}$$

where d_{ij} denotes the distance between zone i and j.

Network association, on the other hand, is defined from the configuration of (abstract) links. Black (1992) defines a binary spatial weight matrix where

two links are considered to be neighbours (weight equals one) if they are (directly) interconnected. We will consider two different spatial weight matrices of this kind. First, we can let the weight equal one for all flows from a zone i (see Fig. 1C). With this definition, we can assess whether all flows from i are small or large, independent of destination j. Hence, using such a spatial weight matrix we can define a measure which is independent of destination j,

$$G_{i\cdot}^{*} = \frac{\sum_j r_{ij} - n\bar{r}}{s\{(t-1)n - n^2/(t-2)\}^{1/2}}, \tag{12}$$

where $\bar{r}$ is the average and s is the standard deviation of flows between all pairs of origins and destinations, and n equals the number of zones[3]. $\bar{r}$ may be approximately equal to zero, for instance if r_{ij} are residual flows from a properly specified regression model.

Second, in a similar way we can define a measure which is independent of origin i by assigning a spatial weight matrix where the weight equals one for all flows with destination in j, as indicated by Fig. 1 D. We define the corresponding measure as

$$G_{\cdot j}^{*} = \frac{\sum_i r_{ij} - n\bar{r}}{s\{(t-1)n - n^2/(t-2)\}^{1/2}}. \tag{13}$$

This measure is independent of origin i and indicates whether the zone j does have large or small inflow.

As defined above, and in Ord and Getis (1995), in the derivation of the G_i statistic it is assumed that all observations are distributed with equal probability under the null hypothesis of no spatial association. This is also the underlying assumption when establishing significance bounds with the conditional permutation approach. However, Bao and Henry (1996) argue that this may be a strong assumption, and they generalise the G_i statistic to allow for spatial heterogeneity. This is relevant also when applied to flow-data. For instance, the observations of flows from zone i may be drawn from a different distribution than the flows from other zones. Consider a origin-destination pair with a significant $G_{ij}(W^d)$ statistic, indicating that there is local nonstationarity in flows from i to a neighbourhood of j. A large $G_{ij}(W^d)$ can also be partially or totally attributed to nonstationarity of generation factors associated with the origin. If this is the case, then the underlying assumption that all flows are drawn from the same distribution may be violated. To control for such errors in generation factors, we can compare the flows from the origin i to a neighbourhood of the destination j with the flows from the origin i only. In equation (7) we let $t = n$ and

$$\bar{r} = \bar{r}_i \triangleq \frac{\sum_{l,l\neq j} r_{il}}{n-1} \tag{14}$$

$$s^2 = s_i^2 \triangleq \frac{\sum_{l,l\neq j}(r_{il} - \bar{r}_i)^2}{n-2}. \tag{15}$$

[3] Note that there is no corresponding $G_{i\cdot}$ statistic, since this statistic would not be independent of j.

We will use the notation $G_{ij}^{i}(W^{d})$ for the measure defined by (7), (9), (14) and (15) to indicate that we compare with flows from i alone, instead of all flows in the sample[4]. In the perspective of spatial heterogeneity of Bao and Henry (1996), the $G_{ij}^{i}(W^{d})$ statistic is derived under the assumption that all flows from (only) zone i are equally probable under the null hypothesis of no spatial association in the flows from zone i alone.

3 First application: Migration model

In this first application we compute the proposed measures to analyse spatial association of migration flows. The migration model used to generate residuals for this application is of standard type for its purpose, viz. a loglinear gravity model estimated with OLS[5]. We estimate gross flows of migrants in east central Sweden, Stockholm county and surrounding regions. There are 53 zones (municipalities), and consequently 2756 OD-pairs[6]. As determinants associated with the origin and destination zones we use: the relative unemployment in the origin zone, the total population in both origins and destinations, ratio between housing prices in origin and destination zones[7], and a dummy-variable if the destination zone is a regional centre. As a determinant describing the distance friction we use distance in kilometers between each pair of zones, and a dummy for origins and destinations within the same county. In Table 1 the estimates are given.

The measures of local spatial association as defined above have been implemented in a computer program, which as input takes spatial weight matrices from a GIS[8] and residual matrices from the estimated migration model. The G_{ij} statistics computed by the program have then been imported into the GIS again for further analysis and visualization. In general, the modelling process can benefit from a close integration with GIS. Since space is always present with localised statistics, a GIS is a natural platform for the analysis. Also, given the amount of data when working with flow data, a GIS is essential for visualization of flow data, even in small applications.

To establish significant bounds we have relied on the conditional permutation approach with 1000 permutations. This approach is recommended in Anselin (1995). The permutation approach as defined in, e.g., Anselin (1995) or Ord and Getis (1995), assumes that all observations (or zones) are equally probable[9]. Hence, under the null hypothesis of no spatial association we assume that the residuals are not correlated. It is important to realize that the residuals do not represent the correct error term in the presence of spatial association. Regression residuals are in fact imperfect estimates for the un-

[4] $G_{ij}^{j}(W^{o})$ statistic can be defined analogously.

[5] Note that a unconstrained gravity model may be seen as a cross product measure of spatial association, as well as a measure of spatial interaction (see Getis 1991). In this perspective, following the arguments of Getis (1995), the residuals from a gravity type model may be considered as data generated from a filtering process.

[6] Data was obtained from Statistics Sweden for the period of 1992–1994.

[7] The ratio between housing prices in origin and destination is included in the model as a determinant of location within a labour market rather than a determinant of migration between labour markets. Housing prices refers to prices of single family dwellings.

[8] We have used a transport oriented GIS, TransCAD.

[9] As discussed earlier, this assumption has been relaxed in Bao and Henry (1996). In the case of $G_{ij}^{i}(W^{d})$ statistic, we randomize over the flows from zone i only, as discussed in Section 2.

Table 1. Estimation results, from migration model. All coefficients are significant at the 99% level

Variable	Coefficient
unemployment	0.628
population, origin	0.736
population, destination	0.680
regional center (dummy)	0.821
housing price ratio	−0.076
distance (km)	−1.071
R^2 adjusted	0.82

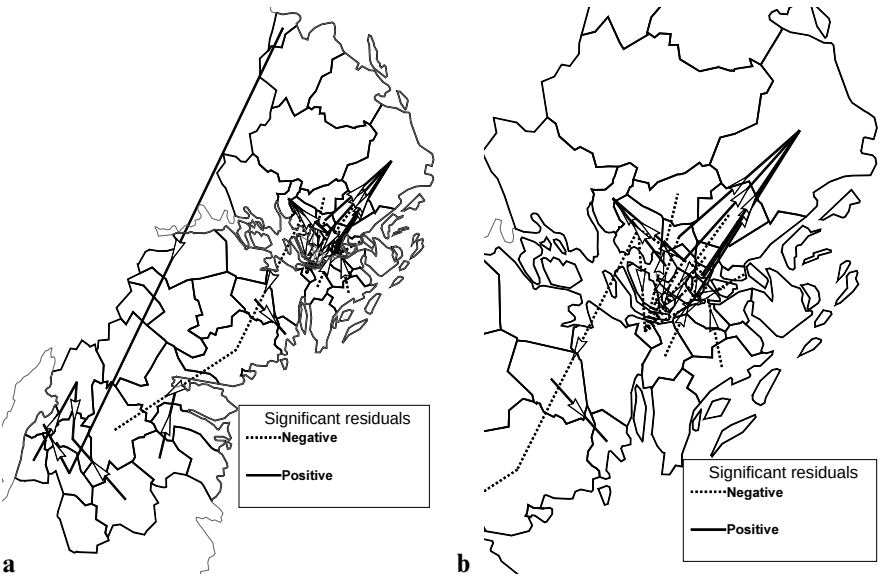

Fig. 2a,b. Flows with high (low) $G_{ij}(W^o)$ statistic, indicating high (low) residual flow from a neighbourhood of zone i to zone j. **(a)** The whole area; **(b)** detail of the Stockholm region

observed error terms in the presence of spatial association[10]. This is not a problem specific for the approach taken in this paper, but rather a general problem when applying e.g. the Moran's I in a residual analysis. However, the problem should be kept in mind when discussing pseudo-significance tests. The G_{ij} statistic applied in this paper should therefore primarily be considered as an explorative tool.

We start our analysis by examining the $G_{ij}(W^o)$ statistic. Intuitively, the statistic identifies pockets of "hot flows" from a neighbourhood of origin i to a destination j.

Considering Fig. 2a,b we observe that some municipalities outside central Stockholm (Salem in the south and Norrtälje in the north) have many significant statistics of the same sign. This indicates that there is spatial non-stationarity in flows when these two municipalities are destinations. Given

[10] See Anselin (1988) pp. 102–103 and references cited there.

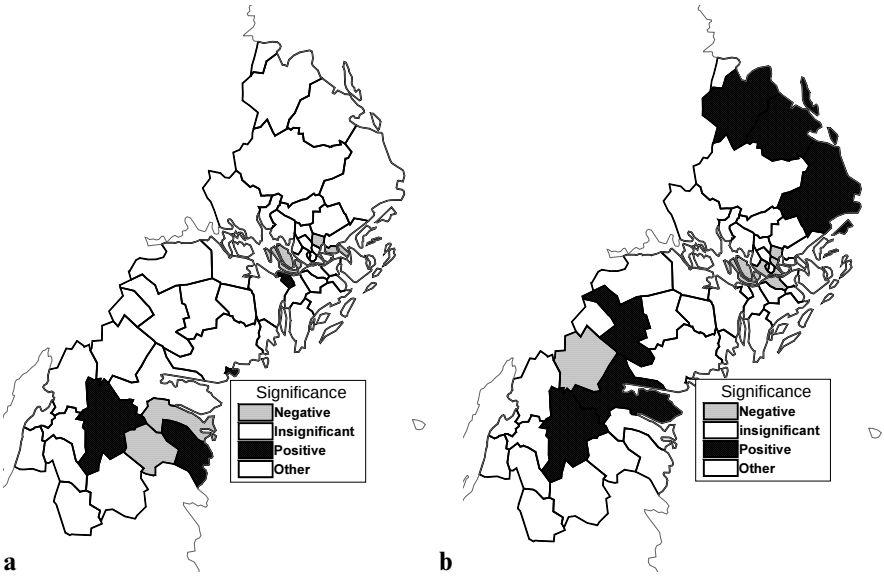

Fig. 3a,b. Zones with significantly high (low) $G_{i\cdot}^*$ statistic, indicating high (low) aggregate outflow from a neighbourhood of zone i and in **b**) zones with high (low) $G_{\cdot j}^*$ statistics indicating high (low) aggregate inflow to a neighbourhood of zone j. **(a)** $G_{i\cdot}^*$; **(b)** $G_{\cdot j}^*$

locations with many significant $G_{ij}(W^o)$, generation or attraction factors may be misspecified. To this end, we apply the $G_{i\cdot}^*$ and $G_{\cdot j}^*$ statistics defined by equations (12) and (13), which are test statistics for local nonstationarities in generation and attraction factors, respectively.

Studying the $G_{i\cdot}^*$ and $G_{\cdot j}^*$ statistics in Fig. 3a and b we find that one of the municipalities with many significantly large $G_{ij}(W^o)$ statistics (Norrtälje) does indeed have significant high $G_{\cdot j}^*$ statistics. This might indicate that the significant $G_{ij}(W^o)$ statistics for this municipality could be attributed to non-stationarity in destination characteristics. Salem on the other hand does not have significant $G_{\cdot j}^*$ statistic, indicating that the negative significant $G_{ij}(W^o)$ statistics can be attributed to origin-destination characteristics. In migration terms, there is a resistance to move to Salem from municipalities in northern Stockholm. This seems plausible given some background knowledge of the socioeconomic pattern in southern and northern Stockholm. The only (proxy) socioeconomic variable included in the model is prices of single family dwellings. The northern parts of Stockholm are characterised by high incomes and high educational level, while the opposite is true for the southern part of Stockholm, such as Salem and its neighbourhood.

Also, note that there are significant $G_{ij}(W^o)$ statistics within the Stockholm region apart from Salem and Norrtälje. This may partly be explained by the structure of the model. The model includes variables reflecting conditions in the labour market, which is standard practice in migration modelling. Only one of the variables (ratio of housing prices) is of relevance for choice of residence. However, the mobility within the Stockholm region is relocation rather than migration.

While the map of $G_{ij}(W^o)$ statistics shows significant values mainly within

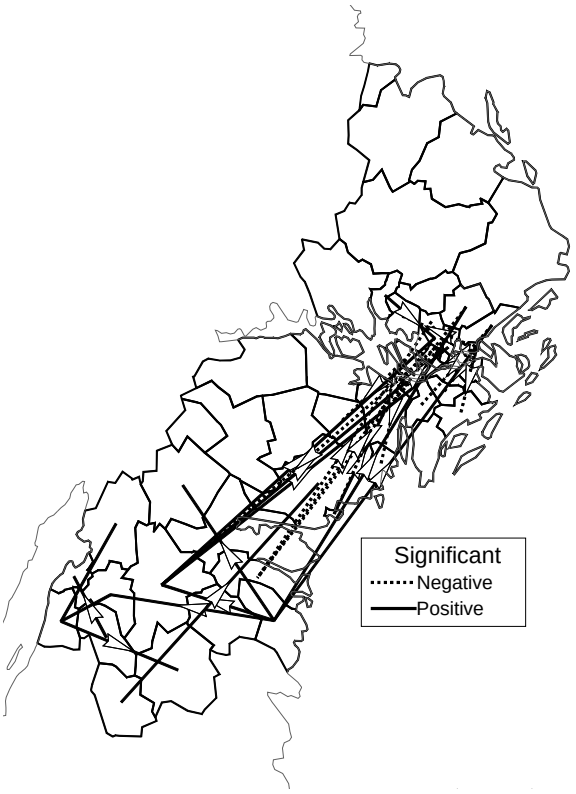

Fig. 4. Flows with high (*low*) $G_{ij}(W^d)$ statistic, indicating high (*low*) residual flow from zone i to a neighbourhood of zone j

the Stockholm region, the $G_{ij}(W^d)$ statistics, considering flows from a zone i to a neighbourhood of zone j, provides us with a quite different picture (see Fig. 4). The pattern of significant $G_{ij}(W^d)$ statistics, have origins in the Stockholm region and destinations in southwest part of the area. The source in the southwest is the university city, Linköping. Also note, in Fig. 3a, that the $G_{i.}^*$ statistic is positively significant for Linköping. However, the $G_{ij}(W^o)$ statistics is not significant at Linköping. This indicates that the non-stationarity indicated by the $G_{ij}(W^d)$ and $G_{i.}^*$ statistics is associated with the origin Linköping itself, and not with the neighbourhood of Linköping. While Linköping is a university city, the surrounding municipalities exhibit quite different characteristics. The analysis shows that in the presence of spatial heterogeneity, like in the neighbourhood of Linköping, it is important to assess nonstationarities with different spatial weight matrices. Studying only the $G_{ij}(W^o)$ statistic there is no evidence of nonstationarity in migration flows from Linköping to Stockholm.

As the $G_{ij}(W^d)$ has been defined, a positive (negative) and significant value indicates that the size of the residuals from a zone i to a neighbourhood of zone j are larger (smaller) than expected. But the significance of the $G_{ij}(W^d)$ may be attributed totally or partially to local nonstationarities in the flows from the origin. In order to test this, we can make use of the measure $G_{ij}^i(W^d)$

as defined by equation (7), (9), (14) and (15). With this measure we compare the flows from the origin i to a neighbourhood of the destination j with the flows from the origin i only, rather than all flows in the sample. This gives a measure which is more sensitive to local spatial association with respect to attraction factors.

As discussed in Sect. 2, the $G_{ij}^{i}(W^d)$ statistics and significance bounds are derived under the null hypothesis of no spatial association in flows from location i. With this measure we control for nonstationarities in generation effects from zone i. With the $G_{ij}(W^d)$ statistic, on the other hand, we do not control for generation effects from zone i. These generation effects, in turn, are assessed by the $G_{i\cdot}^{*}$ statistic. In Fig. 5 the two different statistics are plotted against $G_{i\cdot}^{*}$. As expected, there is some correlation in a), while there is no visible correlation in b)[11]. Correlation of different measures illustrates the importance of considering different spatial weight matrices.

As indicated by Fig. 5, there is a relationship between the $G_{ij}(W^d)$ statistics and the $G_{i\cdot}^{*}$ statistics. We can analyze the stability of the $G_{i\cdot}^{*}$ statistics by studying the variance of $G_{ij}(W^d)$. For instance, note that the variance of the highest $G_{i\cdot}^{*}$ statistic (equal to 6.3) does have a low variance. In fact, the highest $G_{ij}(W^d)$ statistic associated with the zone with the lowest $G_{i\cdot}^{*}$ statistic is just as high as the highest $G_{ij}(W^d)$ statistic associated with zone with the highest $G_{i\cdot}^{*}$ statistic, indicating a large degree of instability.

The relationship between the different statistics can be illustrated by considering an example where all zones have the same number of neighbours, m, e.g. a lattice. Then we can write

$$G_{i\cdot}^{*} = \sum_{j} k_j G_{ij}^{*}(W^d), \tag{16}$$

where

$$k_j = \frac{1}{m} \frac{\{(t-1)m - m^2/(t-2)\}^{1/2}}{\{(t-1)n - n^2/(t-2)\}^{1/2}}. \tag{17}$$

In this case the $G_{i\cdot}^{*}$ statistics can be computed as a (weighted) average of $G_{ij}^{*}(W^d)$. Equation (16) illustrates how the stability of $G_{i\cdot}^{*}$ statistics can be analyzed by the $G_{ij}^{*}(W^d)$ statistics[12]. However, generally the $G_{i\cdot}^{*}$ statistics cannot be written as a weighted sum of $G_{ij}^{*}(W^d)$. The analysis of the stability of the $G_{i\cdot}^{*}$ statistics should therefore be considered as preliminary.

Normality and correlation

Although the G_{ij} statistic is asymptotically normally distributed, in most applications so far the validity of the normal approximation has been unclear. In

[11] Interestingly enough, there is no correlation between $G_{ij}(W^o)$ and $G_{ij}(W^d)$ (not shown in figure). This indicates that the using different spatial weight matrices allow us to study different aspects of local nonstationarity.

[12] This is reminiscent of the definition of LISA (see Anselin 1995). A weighted sum over local indicators of spatial association gives a global measure of spatial association. The $G_{i\cdot}^{*}$ statistic is not a global statistic, but it does remove the flow dimension (direction) in the same way as a global statistic removes space.

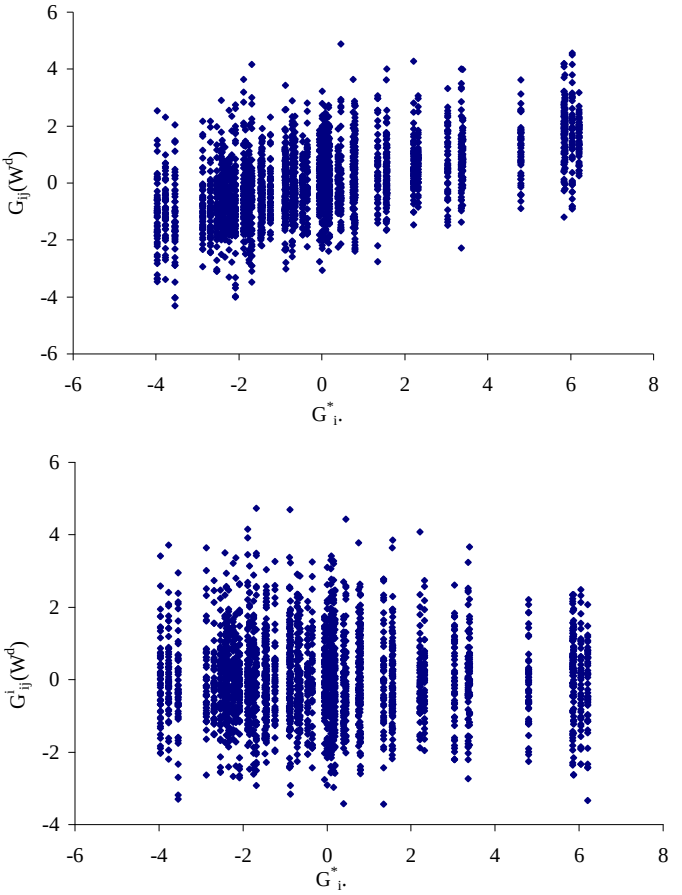

Fig. 5. *Upper* figure $G_{ij}(W^d)$ plotted against $G^*_{i\cdot}$. *Lower* figure G^i_{ij} (W^d) plotted against $G^*_{i\cdot}$

this application with residuals from a OLS model, normality is not a problem, since the residuals from a properly estimated OLS should be normal. To establish this we can examine the distribution of each of the 2756 statistics (the residual matrix contains 2756 elements when the diagonal is excluded). A Jarque-Bera test (JB-test) for each statistic reveals that the null hypothesis of normal distribution cannot be rejected.

However, although being theoretically normally distributed, the G_{ij} statistics are still correlated with each other[13]. This calls for some correction and a permutation approach is appropriate. Since the distribution of the statistic is indeed normal, we can compare an uncorrected normal approximation with the conditional permutation approach. The uncorrected normal assumption

[13] Anselin (1995) discusses more extensively the normal approximation compared to the conditional permutation approach. In the presence of global autocorrelation, the permutation approach is preferable. This holds true also if the underlying distribution (in this example the residuals) is normally distributed.

failed to establish significance in 7 cases when the permutation approach indicated significance, while in 16 cases the normal assumption indicated significance when the permutation approach did not indicate significance, out of the computed 2756 statistics. Whether this is a negligible difference or not depends on the application at hand.

The G_{ij} statistics are easily computed, and usually significant bounds can easily be established with the conditional permutation approach. This is a prerequisite when the statistic is to be used as a method in exploratory spatial data analysis. When working with flow data, the amount of data is often very large. In this application we have 2756 flows. Also, the spatial weight matrix is four dimensional. In a straight forward computation of significance bounds with a general spatial weight matrix, the computational burden becomes large. In this application it took approximately 24 hours to compute significance bounds with the W^{θ} spatial weight matrix, defined in equation (11). This is not feasible in an explorative spatial data analysis. However, if we in an explorative analysis allow for some approximations, the computational burden can be reduced.

For instance, the difference between conditional permutation and permutation with all flows (also the flow from i to j) can be neglected in an application with this amount of flows (and this spatial weight matrix). Furthermore, there is no need to perform permutation for all flows. An average over a small sample of origin-destination pairs provided a good approximation for the significance bounds. The uncorrected normal approximation is also a good approximation in this application with normally distributed statistics. Finally, it is not necessary to perform as much as 1000 permutations. In fact, 50 or 100 permutations also provided good approximations for the significance bounds. However, it is important to further assess the exact implementation of approximations to establish significance bounds with the G_{ij} statistic, and general spatial weight matrices in order for these measures to be feasible in explorative spatial data analysis.

4 Second application: Commuting model

In Sect. 3 we have applied the G_{ij} statistic in a residual analysis of a migration model. The application demonstrated some specific properties of the Getis-Ord statistic applied on residuals from a flow model. In this section we apply the G_{ij} statistic to a small commuting example with only 15 zones. The residuals in this example originate from a logit model of combined mode and destination choice. The model was estimated on the commuting pattern in a region of 15 municipalities in middle Sweden.

In contrast to the application in Sect. 3, the residuals from the commuting model are not theoretically normally distributed. This will enable us to demonstrate the usefulness of the G_{ij} measure of local spatial association in a case where the distribution is not theoretically known. Also, with only 15 zones, small sample properties are important to address.

The model used in this section is a doubly constrained logit model for simultaneous mode and destination choice. This model is of standard type for its purpose. The mode choice set is car and public transport. As determinants we used travel cost for both modes and for public transport time was also

Table 2. Parameter estimates, all coefficients are significant at the 99% level. Parameters for origin zones are left out

Variable	Value
public transport const.	−1.263
travel time publ.	−0.019
travel cost car	−0.033
travel cost publ.	−0.022

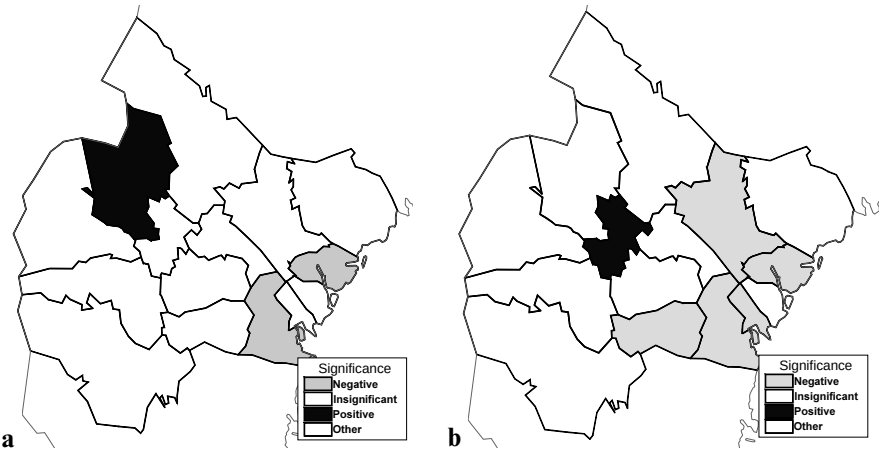

Fig. 6a,b. Zones with high (*low*) $G_{i\cdot}^*$. statistic, indicating high (*low*) aggregate outflow from a neighbourhood of zone i and in **b**) zones with high (*low*) $G_{\cdot j}^*$ statistics indicating high (*low*) aggregate inflow to a neighbourhood of zone j. **(a)** $G_{i\cdot}^*$; **(b)** $G_{\cdot j}^*$

included. Travel cost by car is the marginal travel cost. Travel time is simply based on the shortest path.

In Table 2 the estimates are given. All coefficients have reasonable values which are highly significant.

We have again relied on the permutation approach to establish significance bounds. Note that if we were to apply the G_{ij} statistic to raw flow data with OD-constraints, we would have to adjust the permutation approach using similar formulas as derived by Bao and Henry (1996). However, we assume that the residuals are uncorrelated under the null hypothesis of no spatial association. We have already included the OD-constraints into the model, and the residuals are thus uncorrelated and equal probable under the null hypothesis. Therefore, analysing residuals from our model, we do not have to adjust the permutation approach with respect to OD-constraints.

To begin with, in Fig. 6a the significant $G_{i\cdot}^*$ statistics according to equation (12) are mapped, indicating large and small outflows[14]. In Fig. 6b the corresponding $G_{\cdot j}^*$ statistic is mapped, indicating high and low inflows. In only two zones are both $G_{\cdot j}^*$ and $G_{i\cdot}^*$ small. The zones are both located in the eastern

[14] We will in this section analyse the residuals of car commuting. The residuals of each mode can be analyzed similarly.

Fig. 7a,b. Flows with high (*low*) $G_{ij}(W^d)$ and $G_{ij}(W^o)$ statistic, indicating high (*low*) residual flow from zone i to a neighbourhood of zone j and high (low) residual flow a neighbourhood of zone i to zone j, respectively. **(a)** $G_{ij}(W^d)$; **(b)** $G_{ij}(W^o)$

part of the region. This suggests that this part of the region exhibits more "hot spots". To establish significance bounds we relied upon the conditional permutation approach as described above.

In Fig. 7a and b the significant $G_{ij}(W^d)$ and $G_{ij}(W^o)$ can be seen. Clearly, the zones in the south-east exhibit some significantly large residual flows between themselves, and significantly small flows out to their neighbourhoods. In contrast to the migration application in Sect. 3, in this example the $G_{ij}(W^d)$ statistics and the $G_{ij}(W^o)$ support the same picture of nonstationarities in the eastern part of the area. If there were barrier network effects, rather than nonstationarities among generation or attraction variables, we would expect to find significant statistics of the same sign with both weight matrices[15].

In this application with only 15 zones, small sample issues are important to address. There are really two different issues to be considered. First, when the underlying distribution is not known and the number of neighbours is small compared to the total number of zones, the distribution of the G_{ij} statistic is not known[16]. For most of the measures implemented here, this consideration is relevant. However, in the case of $G^*_{i.}$ statistics (as well as the $G^*_{.j}$ statistic), even for a relatively small number of zones there are many terms to be summed in the numerator of equation (12). Following the arguments of Ord and Getis (1995) a normal approximation may then be more reasonable. To test if the $G^*_{i.}$ and $G^*_{.j}$ statistics are normal, we have used a Jarque-Bera test. The JB statistics of $G^*_{i.}$ and $G^*_{.j}$ was approximately equal to 13 for all 15 zones, the critical value (95%) being 9.21. The JB test on $G_{ij}(W^d)$ and $G_{ij}(W^o)$, on the

[15] In an earlier version of this paper we have used the G_{ij} statistics in an explorative analysis to test for barrier effects.

[16] With respect to normal approximations, Getis and Ord (1992, pp. 191–192) discuss the implications of small total sample size, as well as small number of neighbours.

other hand, strongly rejected the null hypothesis of normally distributed G_{ij} statistics.

Second, with only a small number of zones, we have a small sample of the underlying distribution itself, which is of importance when establishing significance bounds with the conditional permutation approach. In particular, this concerns the G_{ij}^{i} statistic, where we work with only n zones. Since the distribution of residuals from each zone i is non-normal in this application, it is difficult to analyze the $G_{ij}^{i}(W^{d})$ statistics, given unknown small sample properties. However, note that this aspect of small sample properties is of less concern with the G_{ij} statistic applied to flow data. With n zones we have in the magnitude of $n \times n$ flows. This may be a small sample, but considerably less so than what is typically the case in applications of the G_i statistic.

5 Conclusions

The purpose of this paper was to generalise the G_i statistic, put forward by Ord and Getis (1995), to allow for applications with flow-data, and to demonstrate its usefulness in two applications. We have explored nonstationarities and identified underlying geographical patterns. The localised statistics as implemented in this paper makes it possible to address how relationships between variables vary over space[17]. We believe that the used measures have improved our understanding of the strengths and weaknesses of the estimated models in terms of a spatial analysis. This understanding can be incorporated into improved and more comprehensive models.

The application of the G_i statistic to flow data introduces new aspects which merit further consideration on its own. The choice of spatial weight matrix in flow space is one such aspect. Ideally, the spatial weight matrix should be derived from theory[18]. However, in practice it is rarely possible to discriminate among different candidates, and often just a binary spatial weight matrix is used. But with flow data, even the common binary weight matrix is not uniquely defined. In this paper we have defined two different binary spatial weight matrices, one with focus on flows from a zone i to a neighbourhood of zone j, and one with its focus on flows from a neighbourhood of i to a zone j. These two spatial weight matrices focus on different aspects of local nonstationarity. Hence, with use of different spatial weight matrices we have been able not only to identify *where* there is local nonstationarity, but also to some degree the nature of the nonstationarity.

We have also computed the statistic with a more general spatial weight matrix, as proposed by Bolduc (1992) and Bolduc et al. (1995). In this paper, however, we have primarily relied on spatial weight matrices which are binary, defined from the common binary spatial weight matrix. We have found these matrices useful in our explorative analysis, while the more general formulation, equation (11), did not really add to our understanding of the underlying geographical pattern. Generally, the nature of a spatial weight matrix in flow space need to be further explored. Another interesting and important aspect is whether the local statistics with different spatial weight matrices in

[17] See Brunsdon et al. (1996).
[18] See Anselin (1988).

flow space can give us any guidance as to what spatial weight matrix should be incorporated in a more comprehensive and improved model, such as in the modelling approach put forward by Bolduc (1992) and Bolduc et al. (1995). The explorative spatial data analysis used in this paper can in this perspective be seen as a method to identify spatial weight matrices which are best suited to incorporate spatial dependencies in a modelling context.

Another method, Geographical Weighted Regression, for explorative analysis of nonstationarities has been proposed by Brunsdon et al. (1996). With GWR a model is estimated for each observation point in space, and the observations are spatially weighted. This GWR method is applicable to many estimators and models, not only regression models. It is also, in principle, applicable to flow models. The GWR has potential to directly explore how relationships between different sets of variables varies over space[19]. The statistical measures used in this paper assess this only indirectly with different spatial weight matrices. However, given the simplicity of the G_{ij} measure, both intuitively and computationally, we feel that it is an interesting issue to further explore the potential of applying these different methods to flow data.

The G_{ij} statistic in this paper has been implemented in a computer program, and then visualized in a GIS. In the perspective of the growing interest in explorative spatial data analysis (ESDA) methods (see, e.g., Anselin and Bao 1997; Unwin 1996; Ding and Fotheringham 1992), it is a natural next step to implement the G_{ij} statistic more fully integrated with a GIS. In the same way as the G_i statistic has been implemented as macros in GIS software such as ArcInfo (Ding and Fotheringham 1992) and ArcView (Scott and Lloyd 1997), we think it would be worthwhile to implement the G_{ij} statistical measures of local spatial association in transportation related GIS, such as TransCAD. Although the statistic G_{ij} and G_i are formally equivalent, the special characteristics introduced with flow data, compared with only scalar data, motivate special consideration.

In this paper we have employed the G_i statistic, but also local Moran's I can be computed in a similar manner. As stated by Ord and Getis (1995) and Anselin (1995), both statistics have their advantages, and both should be calculated in any exercise analyzing spatial association. The advantages of each measure in a flow-data context should also be further studied.

References

Anselin L (1988) *Spatial econometrics: Methods and models.* Kluwer, Dordrecht, The Netherlands

Anselin L (1995) Local indicators of spatial association – LISA. *Geographical Analysis* 27(2):93–115

Anselin L, Bao S (1997) Exploratory spatial data analysis linking SpaceStat with arcview. In: Fischer MM, Getis A (eds) *Recent developments in spatial analysis* (Advances in Spatial Science). Springer, Berlin Heidelberg New York, pp 35–59

Bao S, Henry M (1996) Heterogeneity issues in local measurements of spatial association. *Geographical Systems* 3(1):1–31

Besag J, Newell J (1991) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society* (A) 154:143–155

[19] In principle, the GWR method result in one set of parameter estimates for each origin-destination pair.

Black WR (1992) Network autocorrelation in transport network and flow systems. *Geographical Analysis* 24(3):207–222

Bolduc D (1992) Spatial autoregressive error components in travel flow models. *Regional Science and Urban Economics* 22(3):371–385

Bolduc D, Laferriere R, Santarossa G (1995) *Spatial autoregressive error components in travel flow models: An application to aggregate mode choice.* In: Anselin L, Florax RJGM (eds) New directions in spatial econometrics (Advances in Spatial Science). Springer, Berlin Heidelberg New York, pp 96–108

Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis* 28(4):281–298

Ding Y, Fotheringham AS (1992) The integration of spatial analysis and GIS. *Computers, Environment and Urban Systems* 16:3–19

Fotheringham, AS (1992) Explorative spatial data analysis and GIS. *Environment and Planning* (A) 25:156–158

Fotheringham AS (1994) On the future of spatial analysis: The role of GIS. *Environment and Planning* (A) Anniversary Issue, 30–34

Fotheringham AS, Charlton M, Brunsdon C (1997) *Measuring spatial variations in relationships with geographical weighted regression.* In: Fischer MM, Getis A (eds) Recent developments in spatial analysis (Advances in Spatial Science). Springer, Berlin, Heidelberg New York, pp 60–82

Getis A (1991) Spatial interaction and spatial autocorrelation: a cross product approach. *Environment and Planning* (A) 23(9):1269–1277

Getis A (1995) *Spatial filtering in a regression framework: Examples using data on urban crime, regional inequality, and government expenditures.* In: Anselin L, Florax RJGM (eds) New Directions in Spatial Econometrics (Advances in Spatial Science). Springer, Berlin Heidelberg New York, pp 172–185

Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24(3):189–206

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27(4):286–305

Scott LM, Lloyd WJ (1997) *Spatial analysis in a GIS environment: Employment patterns in greater Los Angeles, 1980–1990.* Proceedings of the 1997 UCGIS annual assembly and summer retreat

Unwin A (1996) Exploratory spatial data analysis and local statistics. *Computational Statistics* 11:387–400